UNC
ESHELMAN
SCHOOL OF PHARMACY

MML
UNC.EDU

QSAR Modeling on the Web.

ChemBench:
Free Online QSAR Modeling Tool

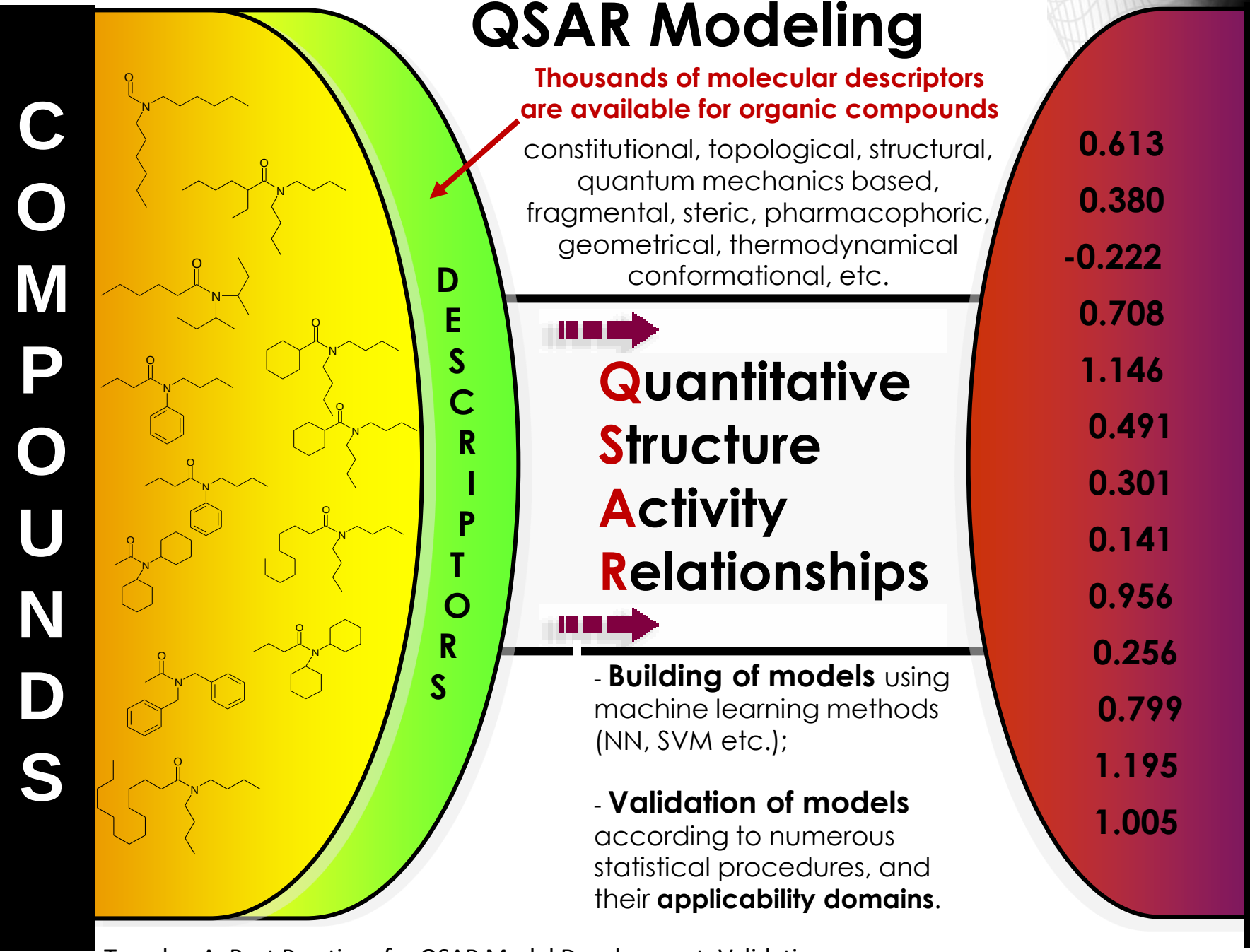
Diane Pozefsky, Diptorub Deb, Chi Xie, Alexander Sedykh, Alexander Tropsha

Laboratory for Molecular Modeling
Eshelman School of Pharmacy, UNC-Chapel Hill

Brief overview of QS[A,P,N,T]R progression

- Experimental Data = pain
  - Structure
  - Activity
- Validated models of data
  - Descriptors
  - Statistical/machine learning techniques
- Imputed data
- Experimentally confirmed predictions
- Reliable models to enable decision support (both in research and regulations) = gain

QSAR Modeling



Thousands of molecular descriptors are available for organic compounds

constitutional, topological, structural, quantum mechanics based, fragmental, steric, pharmacophoric, geometrical, thermodynamical conformational, etc.

Quantitative Structure Activity Relationships

- **Building of models** using machine learning methods (NN, SVM etc.);

- **Validation of models** according to numerous statistical procedures, and their **applicability domains**.

Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation Mol. Inf., 2010, 29, 476 – 488

Published guidance on model development and validation: J. Dearden's 21 "how not to do QSAR" principles

Table 1. Types of error in QSAR/QSPR development and use.

No.	*Type of error*	*Relevant OECD principle(s)*
1	Failure to take account of data heterogeneity	1
2	Use of inappropriate endpoint data	1
3	Use of collinear descriptors	2, 4, 5
4	Use of incomprehensible descriptors	2, 5
5	Error in descriptor values	2
6	Poor transferability of QSAR/QSPR	2
7	Inadequate/undefined applicability domain	3
8	Unacknowledged omission of data points	3
9	Use of inadequate data	3
10	Replication of compounds in dataset	3
11	Too narrow a range of endpoint values	3
12	Over-fitting of data	4
13	Use of excessive numbers of descriptors in a QSAR/QSPR	4
14	Lack of/inadequate statistics	4
15	Incorrect calculation	4
16	Lack of descriptor auto-scaling	4
17	Misuse/misinterpretation of statistics	4
18	No consideration of distribution of residuals	4
19	Inadequate training/test set selection	4
20	Failure to validate a QSAR/QSPR correctly	4
21	Lack of mechanistic interpretation	5

Dearden JC et al., **2009,** SAR and QSAR in Environmental Research, Vol. 20, Nos. 3–4, April–June 2009, 241–266.

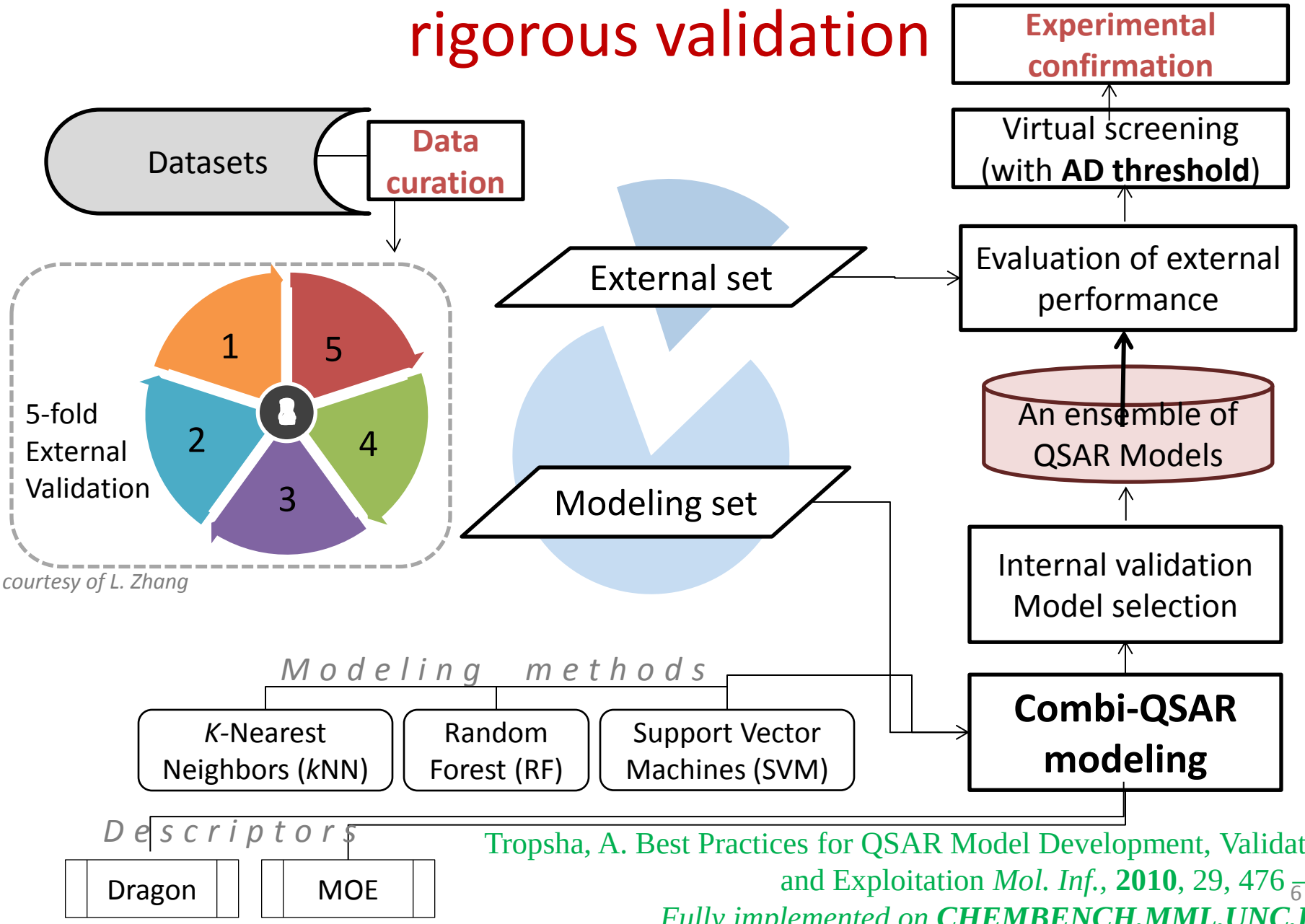
The OECD Principles of Model Validation*

Fully implemented within our modeling workflow and within ChemBench, chembench.mml.unc.edu

1. A defined endpoint
2. An unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of-fit, robustness and predictivity
5. A mechanistic interpretation
6. *Proposed: Chemical structures should be curated and harmonized (should be added!)*

*http://www.oecd.org/dataoecd/33/37/37849783.pdf

QSAR Modeling Workflow: the importance of rigorous validation



Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation *Mol. Inf.*, **2010**, 29, 476 – 488

Fully implemented on ***CHEMBENCH.MML.UNC.EDU***

Data dependency and data quality are

MML UNC.EDU

ChemBark

News, Analysis, and Commentary for the World of Chemistry & Chemical Research

« Hacks for Septa

Organometallics Responds to the Dorta Situation »

A Disturbing Note in a Recent SI File

August 6th, 2013

A recently published ASAP article in the journal *Organometallics* is sure to raise some eyebrows in the chemical community. While the paper itself is a straightforward study of palladium and platinum bis-sulfoxide complexes, page 12 of the corresponding Supporting Information file contains what appears to be an editorial note that was inadvertently left in the published document:

> Emma, please insert NMR data here! where are they? and for this compound, just make up an elemental analysis...

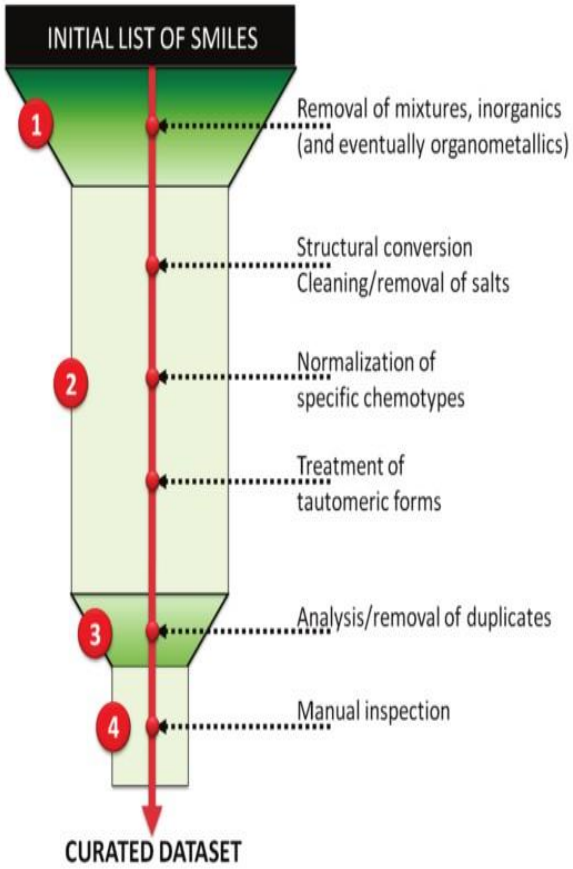
This statement goes beyond a simple embarrassing failure to properly edit the manuscript, as it appears the first author is being instructed to fabricate data. Elemental analyses would be very easy to fabricate, and long-time readers of this blog will recall how fake elemental analyses were pivotal to Bengu Sezen's campaign of fraud in the work she published from 2002 to 2005 out of Dalibor Sames' lab at Columbia.

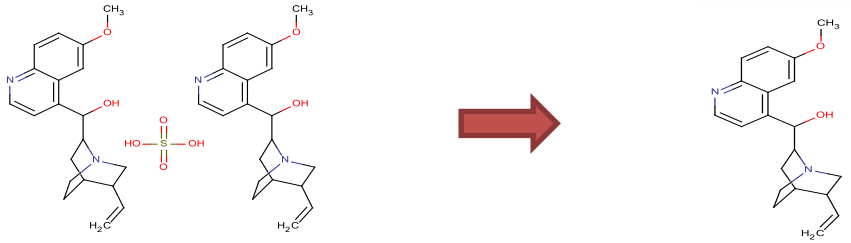
The compound labeled **14** (an acac complex) in the main paper does not appear to correspond to compound **14** in the SI. In fact, the bridged-dichloride compound appears to be listed an as unlabeled intermediate in Scheme 5, which should raise more eyebrows. Did the authors unlist the compound in order to avoid having to provide robust characterization for it?

ChemBark is contacting the corresponding author for comment, and his response will be posted in full when we receive it.

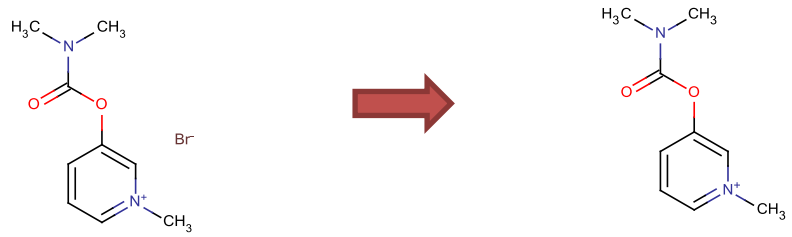
Chemical Structure Curation

Chemical structures should be cleaned and standardized (duplicates removed, salts stripped, neutral form, canonical tautomer, etc) to enable rigorous model development

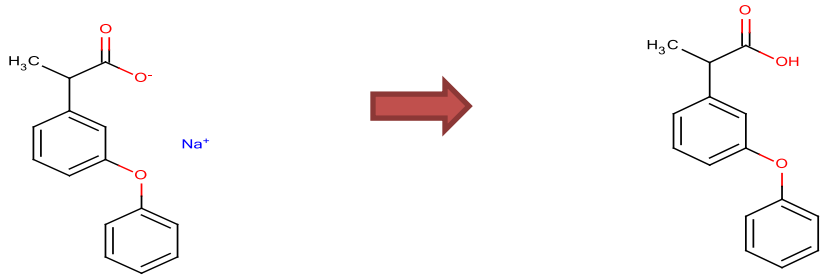




- Quinine sulfate dihydrate



- Pyridostigmine Bromide

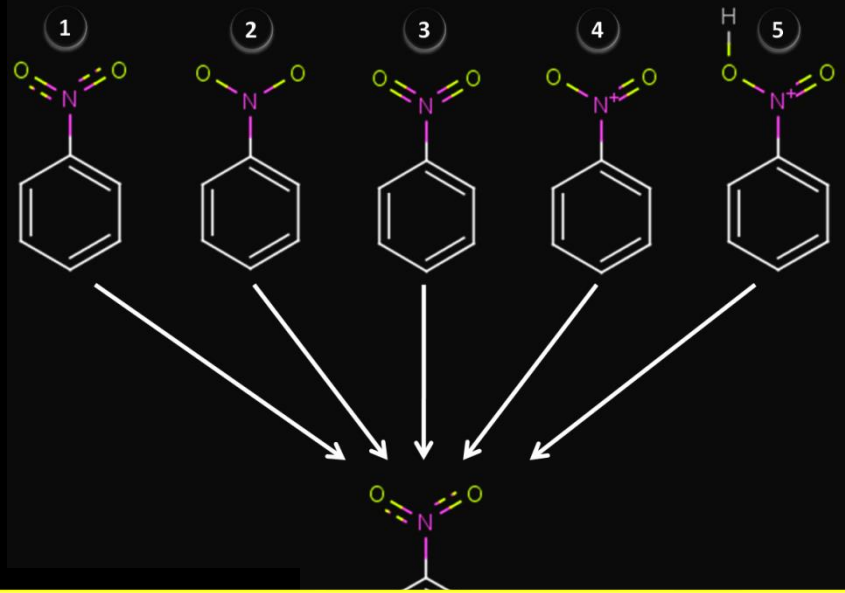


- Fenoprofen Sodium

Muratov, Fourches, Tropsha. Trust but verify.
JCIM, 2010, 29, 476 – 488.

QSAR modeling of nitro-aromatic toxicants

-Case Study 1: 28 compounds tested in rats, log(LD50), mmol/kg.
-Case Study 2: 95 compounds tested against *Tetrahymena pyriformis,* log(IGC50), mmol/ml.



Data curation affects the accuracy (up or down!) of QSAR models

Even small differences in structure representation can lead to significant errors in prediction accuracy of models

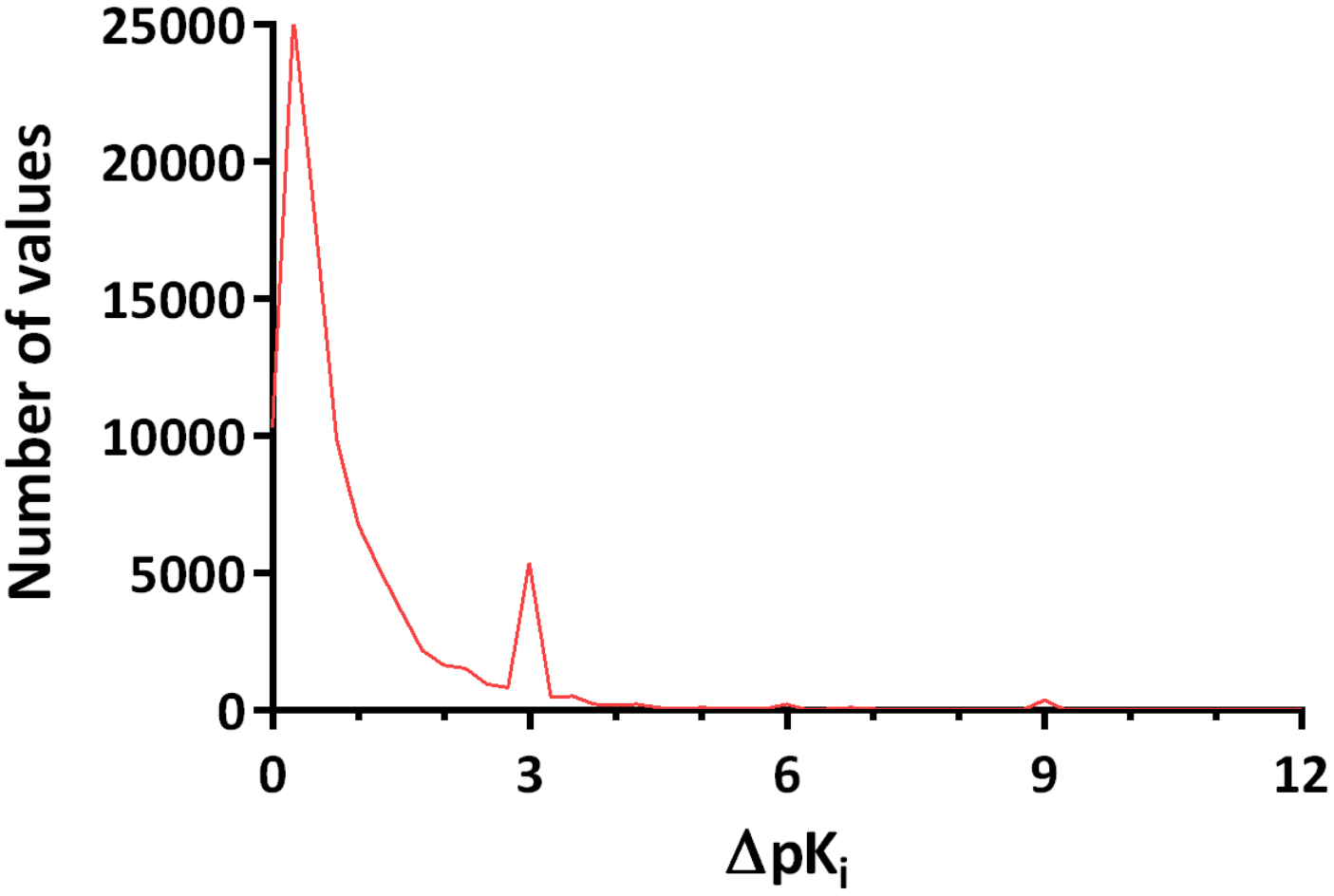
Artemenko, Muratov et al. J. SAR QSAR 2011, 22 (5-6), 1-27.

Manual Curation of the ChEMBL database (following several automated steps)

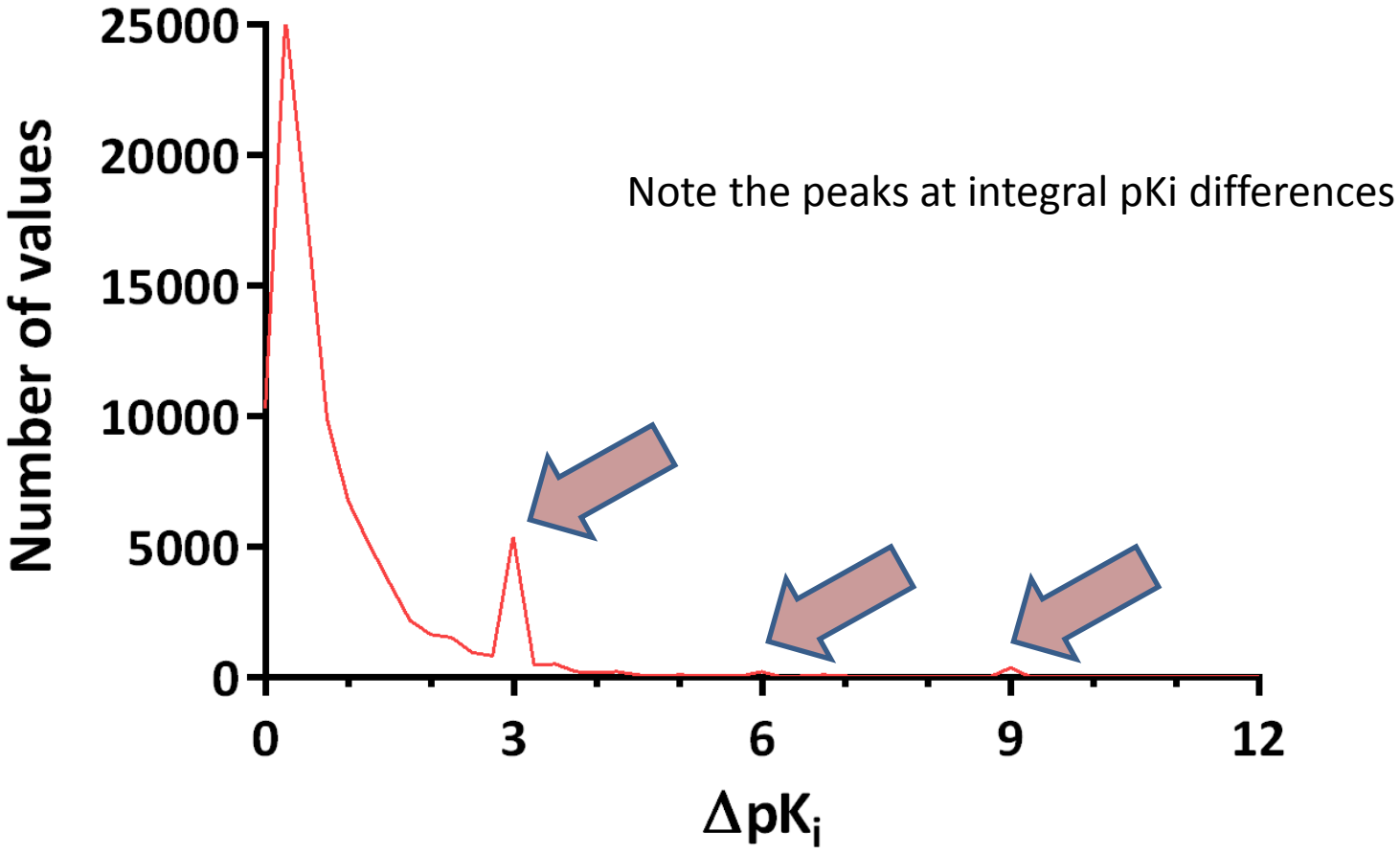
- Input: 190,068 compound-target measures in pairs of papers
  - Used values as published in ChEMBL
  - Converted to standardized pK_i values
  - Semi-automated (based on units and type of value reported)
- 23,956 failed to be automatically converted
  - Mostly Log K_i or –Log K_i values but others
  - Manually examined papers representing ~70% and hand converted affinity value, except when data was being recycled/recited
- Final: 178,317 total replicate pairs of values

Fant et al, manuscript in preparation

Frequency distribution plot for differences in pK_i values (>1%) for duplicates



A Recurrent Pattern



Cheminformatics Analysis of qHTS data

over 17,000 compounds screened against five major CYP isozymes using In Vitro bioluminescent qHTS assay

	#	SID	CID	CID (TXT FILE)	Nition Obse	2c19_LogAC50	2d6_LogAC50	3a4_LogAC50	1a2_LogAC50	2c9_LogAC50	Compound QC
51	7955	11113498	1348	1348	TRUE	-6.1	-5.7	-5.1	-5.9	-5.4	QC'd by Tocris
60	7577	11113881	1370	1370	TRUE	-4.9	-5	-4.8	-5.6	-5.1	QC'd by Tocris
69	7888	11113566	1574	1574	TRUE	-5.1	-4.7	-4.8	-4.7	-4.4	QC'd by Tocris
97	7686	11113772	1797	1797	TRUE	-5	-4.6	-4.4	-7.4	-4.6	QC'd by Tocris
117	7987	11113466	1960	1960	TRUE	-5.2	-4.6	-4.8	-4.8	-4.6	QC'd by Tocris
130	7925	11113529	2052	2052	TRUE	-4.8	-4.7	-4.5	-5.3	-5.1	QC'd by SigmaAldrich
136	7531	11113928	2125	2125	TRUE	-5.1	-5.4	-5	-4.8	-5.7	QC'd by Tocris
210	9989	11110929	2703	2703	TRUE	-5	-4.6	-4.5	-5	-4.4	QC'd by SigmaAldrich
227	9973	11110952	2782		TRUE	-6.7	-5.9	-5.2	-5	-4.6	QC'd by SigmaAldrich
229	7772	11113684	2790	2790	TRUE	-4.8	-4.9	-5.8	-4.8	-4.9	QC'd by Tocris
240	9964	11110963	2812	2812	TRUE	-5.1	-5	-7.3	-5.4	-6.5	QC'd by Prestwick
241	9965	11110962	2812		TRUE	-5	-4.4	-6.9	-4.8	-6	QC'd by SigmaAldrich
242	8112	11113341	2818	2818	TRUE	-4.6	-4.8	-4.5	-4.8	-4.4	QC'd by Tocris
264	9208	11111961	2998	2998	TRUE	-5.1	-4.6	-5.4	-4.9	-5.5	QC'd by SigmaAldrich
282	7920	11113534	3101	3101	TRUE	-7.2	-6.1	-5.5	-7.7	-7	QC'd by Tocris
283	9889	11111058	3101		TRUE	-6.3	-5.4	-5.5	-6.9	-6	QC'd by SigmaAldrich
290	9873	11111076	3136	3136	TRUE	-4.5	-4.4	-4.7	-5.4	-4.4	QC'd by SigmaAldrich
309	8948	11112239	3293	3293	TRUE	-7.3	-5.6	-4.9	-5.3	-5.7	QC'd by Prestwick
326	9809	11111163	3396		TRUE	-4.8	-5	-5.2	-4.9	-4.4	QC'd by SigmaAldrich
345	7961	11113492	3455	3455	TRUE	-4.6	-6.2	-4.9	-4.5	-4.7	QC'd by Tocris
353	8100	11113353	3488	3488	TRUE	-5	-5	-5	-4.4	-5.1	QC'd by Tocris
364	7374	11114090	3538	3538	TRUE	-5.1	-4.6	-5.3	-4.5	-5.9	QC'd by Tocris
383	7284	11114182	3671	3671	TRUE	-5.5	-7.4	-5.1	-6.2	-6.2	QC'd by SigmaAldrich
384	9442	11111654	3675	3675	TRUE	-6.5	-5.6	-5.1	-6	-6.8	QC'd by Prestwick
385	9443	11111653	3675		TRUE	-6.1	-5.2	-5.5	-5.5	-5	QC'd by SigmaAldrich
394	8391	11112811	3698	3698	TRUE	-5.3	-4.9	-5.5	-4.8	-4.9	QC'd by Prestwick
410	9189	11111983	3797		TRUE	-4.5	-5.7	-5.7	-5.4	-4.9	QC'd by SigmaAldrich
422	9652	11111370	3885	3885	TRUE	-5.4	-4.8	-4.8	-5.4	-4.5	QC'd by SigmaAldrich
428	7207	11114259	3932	3932	TRUE	-6.7	-5.1	-6.3	-4.5	-5.1	QC'd by SigmaAldrich
485	7988	11113465	4299	4299	TRUE	-8.6	-4.5	-4.6	-4.4	-5.7	QC'd by Tocris
486	7984	11113469	4306	4306	TRUE	-7.4	-5.1	-4.9	-5.6	-4.9	QC'd by Tocris

Nature Biotechnology, 2009,

J. Chem. Inf. Model., 2011,

Duplicate analysis

- Carried out by ISIDA/Duplicates program
- 1,280 duplicate couples were found
  - 406 had a complete matching profile
  - 874 had profile differences
  - A total of 1,535 discrepancies were found in the 874 duplicates couples CYP annotation:

	CYP2C9	CYP1A2	CYP3A4	CYP2D6	CYP2C19
# of discrepancies	**154**	**363**	**426**	**422**	**170**

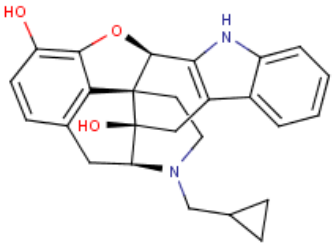
PROBLEM: CYP bioprofiles for some duplicates are dramatically different

➡ Need biological curation!

Neighborhood analysis helps to choose correct value

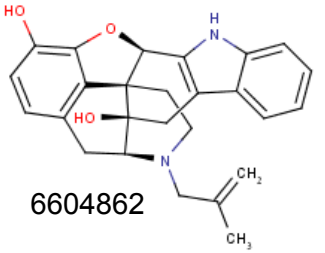
Case Study: structural duplicates found in NCGC CYP450 qHTS data

Tocris-0740	SID	Supplier	Cytochrome P450 2C9	1A2	3A4	2D6	2C19
CID_6603937	11113673	Tocris	-4.6	-4.4	**-4.6**	-6.2	-4.5
CID_6603937	11111504	Sigma Aldrich	-4.4	INA	**-8**	-5.6	-5

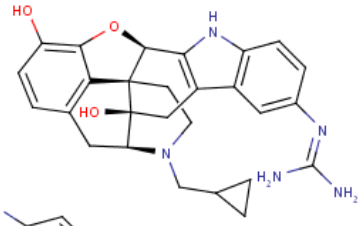


FALSE-POSITIVE

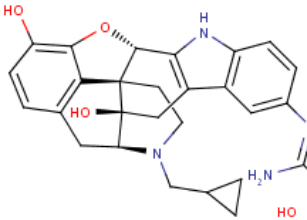
5 Nearest neighbors	Tanimoto Similarity	SID	Supplier	Cytochrome P450 2C9	1A2	3A4	2D6	2C19
6604862	**0.98**	11114071	Tocris	INA	INA	**-4.5**	INA	-5.5
6604106	**0.98**	11112029	Sigma Aldrich	INA	INA	**-5.1**	INA	INA
6604846	**0.98**	11114012	Tocris	INA	INA	**INA**	INA	INA
6604136	**0.95**	11112054	Sigma Aldrich	INA	INA	**-4.8**	-5.9	INA
6604137	**0.95**	11113764	Tocris	INA	-4.4	**-4.7**	-4.5	INA



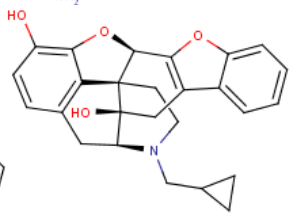
6604862



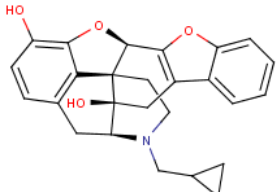
6604106



6604846



6604136



6604137

Examples of Commercial Toxicity Predictors

Prediction tool	Categories of endpoints[a]	Features
ADMET Predictor www.simulations-plus.com	Irritation and adverse health effects Carcinogenicity and genotoxicity Acute and developmental toxicity Endocrine disruption and ecotoxicity	QSAR
ACD/Tox Suite www.acdlabs.com	Irritation and adverse health effects Genotoxicity Acute toxicity Endocrine disruption and ecotoxicity	Confidence intervals and probability of predictions
DEREK, DEREK Nexus www.lhasalimited.org	Irritation and adverse health effects Carcinogenicity and genotoxicity Developmental toxicity	Expert system
TOPKAT www.accelrys.com	Irritation Carcinogenicity and genotoxicity Acute, chronic, and developmental toxicity Ecotoxicity	QSAR
CASE www.multicase.com	Irritation and adverse health effects Carcinogenicity and genotoxicity Acute and developmental toxicity Endocrine disruption and ecotoxicity	Fragment-based QSAR
Leadscope Model Applier www.leadscope.com	Adverse health effects Carcinogenicity Reproductive and developmental toxicity	QSAR
HazardExpertPro, ToxAlert www.compudrug.com	Adverse health effects Carcinogenicity and genotoxicity Developmental toxicity	Expert system

Challenges with using most of the QSAR tools

- Most are commercial; training sets are hidden; very few available online

- Most make binary predictions ("is my compound likely to be mutagenic?" yes/no; Few continuous (produce a number rather than a class) predictors are available (most for LD_{50}, LC_{50}, etc.)

- Most predictors are of a "black box" variety (not transparent)

- Typically, don't consider "domain of applicability "

Examples of Toxicity Predictors in Public Domain

Prediction tool	Categories of endpoints	Features
T.E.S.T. (EPA) www.epa.gov/nrmrl/std/cppb/qsar	Carcinogenicity and genotoxicity Acute and developmental toxicity Ecotoxicity	Consensus and batch prediction modes by QSAR
OncoLogic (EPA) http://www.epa.gov/oppt/sf/pubs/oncologic.htm	Carcinogenicity	Expert system
OpenTox www.opentox.org	Irritation Carcinogenicity and genotoxicity	Expert system (ToxTree); QSAR (Lazar); ontology of toxic endpoints
OECD QSAR Toolbox www.qsartoolbox.org	Irritation Carcinogenicity and genotoxicity Ecotoxicity	Prediction by ''read across'' analysis or by QSAR
OCHEM www.ochem.eu	Genotoxicity Ecotoxicity	Online chemical database and QSAR modeling environment
ChemBench chembench.mml.unc.edu	Genotoxicity Ecotoxicity	Web-based platform for QSAR modeling or prediction

Reviewed in:

Rusyn *et al* Toxicological Sciences 127(1), 1–9 (2012)
"Predictive Modeling of Chemical Hazard by Integrating Numerical Descriptors of Chemical Structures and Short-term Toxicity Assay Data"

CHEMBENCH

http://chembench.mml.unc.edu

HOME | MY BENCH | DATASET | MODELING | PREDICTION

ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

Chembench is a free portal that enables researchers to mine available chemical and biological data. Chembench can help researchers rationally design or select new compounds or compound libraries with significantly enhanced hit rates in screening experiments.

It provides cheminformatics research support to molecular modelers, medicinal chemists and quantitative biologists by integrating robust model builders, property and activity predictors, virtual libraries of available chemicals with predicted biological and drug-like properties, and special tools for chemical library design. Chembench was initially developed to support researchers in the **Molecular Libraries Probe Production Centers Network (MLPCN)** and the Chemical Synthesis Centers.

Please cite this website using the following URL: **http://chembench.mml.unc.edu**

The Carolina Cheminformatics Workbench (Chembench) is developed by the Carolina Exploratory Center for Cheminformatics Research (CECCR) with the support of the **National Institutes of Health** (grants **P20HG003898** and **R01GM066940**) and the Environmental Protection Agency (RD83382501 and RD832720). This website has been

Please login

Username:

Password: login

Or, **login as a guest**

Forget your password? **click here**

New Users

Please **register here**

Help & Links

Chembench Overview
Chembench Workflows & Methodology
Links to More Cheminformatics Tools

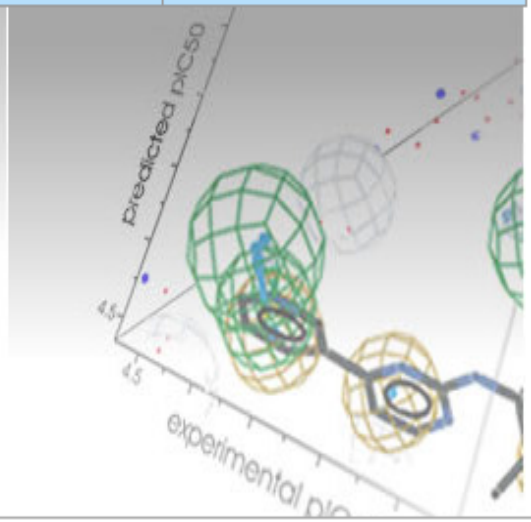
Statistics

Visitors: 350266
Users: 652
Jobs completed: 21130
Compute time used: 25.378 years
Current Users: 1
Running Jobs: 4

My Bench

CHEM BENCH

HOME | MY BENCH | DATASET | MODELING | PREDICTION



My Bench

Every dataset, predictor, and prediction you have created on Chembench is available on this page. You can track progress of all the running jobs using the job queue.

Publicly available datasets and predictors are also displayed. If you wish to share datasets or predictors you have developed with the Chembench community, please contact us at ceccr@email.unc.edu.
All data is sorted by the creation date in descending order (newest on top).

Job Queue | Datasets | Predictors | Predictions

Job Queue

Running jobs from all Chembench users are displayed below. Use the REFRESH STATUS button to update the list. Other users [illegible]e your jobs while they are running, but only you can access your completed datasets, predictors, and predictions.

REFRESH STATUS

Job Queue | Datasets | Predictors | Predictions

Datasets

* Descriptors for the dataset were created outside of Chembench and uploaded by the user.

Click on the name of dataset to visualize it.
*You are currently viewing all available public datasets. You can choose to hide these from the **edit profile** page.*

Name	Number of Compounds	Type	Structure Images Available	Descriptor Type name	Date Created	Public/Private
BCRPi10_3	395	CATEGORY	YES	CDK DRAGONH DRAGONNOH	2013-08-05 11:46	Private

Job Queue | Datasets | Predictors | Predictions

Predictors

* Predictor was built on a dataset with descriptors that were created outside of Chembench and uploaded by the user.

Click on the name of a predictor to analyze the modeling results.

Name	Dataset	Nfold	Act type	External Set R^2 or CCR	Modeling Method	Descriptor Type	Public/Private	Date Created
ASBTi10d_CDK_SA	ASBTi10d	YES	CATEGORY	0.862 ± 0.048	KNN-SA	CDK	Private	2013-08-05 12:26

Job Queue | Datasets | Predictors | Predictions

Predictions

Click on the name of a prediction to see the results.

Name	Dataset	Predictor	Date Created
qqqq	HDAC_59	PPB_lnKa_RandomForest	2013-08-22 09:33

Upload Dataset

CHEM BENCH

Logged in as **atropsha**.
log out | edit profile | help pages

HOME | MY BENCH | DATASET | MODELING | PREDICTION

Upload Dataset Files

Select the type of dataset to create.
For the "Modeling Set" and "Prediction Set", you do not need to provide descriptors; Chembench will generate descriptors as needed for visualization, modeling, and prediction.
For the "Modeling Set With Descriptors" and "Prediction Set With Descriptors", you will need to upload an X file containing the descriptor values.

Modeling Set | Prediction Set | Modeling Set With Descriptors | Prediction Set With Descriptors

Prediction Dataset

A dataset will be created from the SDF file you supply.

SDF File: Browse...

Standardize structures: ☑

Generate M-heatmap: ☑ *(Unchecking this box will accelerate dataset generation but will eliminate heatmap based on Mahalanobis distance measure)*

Define External Validation Scheme

Logged in as **atropsha**.
log out | edit profile | help pages

HOME | MY BENCH | DATASET | MODELING | PREDICTION

Define External Set

A subset of the compounds in the dataset will be reserved for testing of the models you build. If you already have a test set defined, use the "Choose Compounds" tab to pick those compounds as your external test set.
These parameters only apply to modeling sets.

Random Split | Choose Compounds | n-Fold Split

Set Automatic Splitting Parameters

Use activity binning: ☑
External Set Size: 20 Percent

Create Dataset

A job will be started to generate visualizations and chemical sketches for this dataset.

Dataset Name:
Reference (optional):
Description (optional):

Create Dataset

Build Predictor (Model)

Logged in as jwignall@unc.edu.
log out | edit profile | help pages

HOME | MY BENCH | DATASET | MODELING | PREDICTION

Chembench Model Development

Select Descriptors

Descriptor Type:

- (•) CDK [202 descriptors]
- ( ) MolconnZ [375 descriptors]

Choose Model Generation Method

Random Forest | Support Vector Machines | GA-kNN | SA-kNN

Set Random Forest Parameters

Choose Internal Data Splitting Method

Random Split | Sphere Exclusion

Set Random Splitting Parameters

Select Predictor(s)

HOME | MY BENCH | DATASET | MODELING | PREDICTION

Select Predictors | Compounds

+ Drug Discovery Predictors

+ ADME Predictors

- Toxicity Predictors

Select	Name	ACTIVITY	TYPE	TRAIN/TEST	ACCURACY*
☐	5HT2B_Binder_Dr	Acute toxicity, rat	*category*	295/74	0.80-0.82
☐	Ames_Genotoxi	Acute toxicity, rat	*continuous*	3472/3913	0.24-0.70
☐	Ames_Genotoxi	Genotoxicity	*category*	~4500/2000	~0.85
☐	RAT_ACUTE_	ER-alpha binding	*continuous*	437/109	0.73
☐	T.Pyriform	ER-beta binding	*continuous*	110/27	0.53
		MDR1 transport	*category*	435/109	0.76
		Aquatic toxicity	*continuous*	644/449	0.67-0.85
		Skin sensitization	*category*	210/52	0.75-0.77
		5HT2B binding	*category*	243/79	0.8
		Blood-brain barrier	*continuous*	144/381	0.59-0.80
		Plasma protein binding	*continuous*	995/422	0.66-0.68

- Private Predictors

Select	Name
☐	ASBTi10d_CDK_

Select Dataset for Prediction

HOME | MY BENCH | DATASET | MODELING | PREDICTION

Select Predictors | Compounds

Chosen Predictors:

Name	Date Created	Modeling Method	Descriptor Type
5HT2B_Binder_DragonkNN	2010-09-16 03:57	KNN	DRAGONH

Select a Dataset

Select a Dataset: ACE-benchmark [View Dataset]

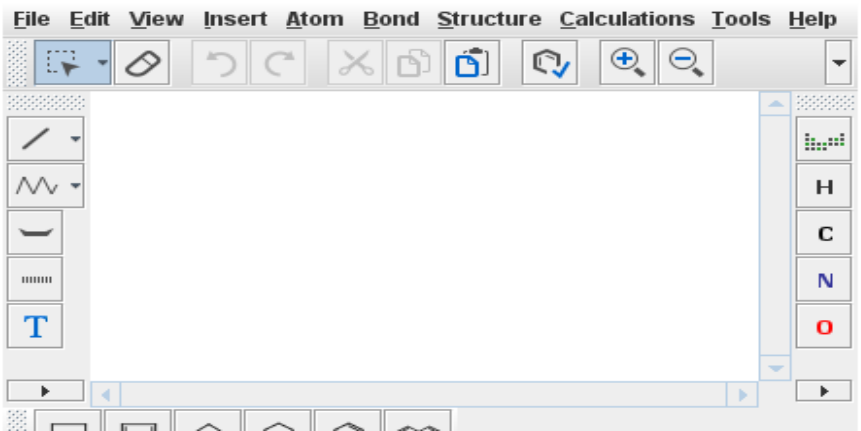
(Use the "DATASET" page to create datasets.)

Applicability Cut Off: Do not use

Prediction Name:

[Submit Prediction Job]

Or Enter a Compound (Sketch a Compound Or Enter a SMILES String)



Enter a molecule in SMILES format, e.g. C1=CC=C(C=C1)CC(C(=O)O)N (phenylalanine). Or, use the applet on the left to draw a molecule, then click "Get SMILES and Predict".

Note: If the sketch applet did not load, your Java version may be out of date. You can download an updated version **here**.

SMILES:

Applicability Cut Off: Do not use [Predict]

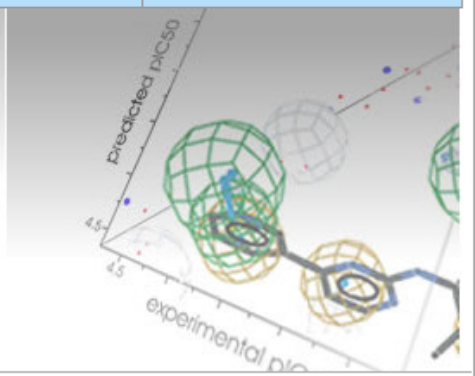
Check Prediction Job Status

HOME | MY BENCH | DATASET | MODELING | PREDICTION

My Bench

Every dataset, predictor, and prediction you have created on Chembench is available on this page. You can track progress of all the running jobs using the job queue.

Publicly available datasets and predictors are also displayed. If you wish to share datasets or predictors you have developed with the Chembench community, please contact us at ceccr@email.unc.edu.
All data is sorted by the creation date in descending order (newest on top).



Job Queue | Datasets | Predictors | Predictions

Job Queue

Running jobs from all Chembench users are displayed below. Use the REFRESH STATUS button to update the list. Other users can see your jobs while they are running, but only you can access your completed datasets, predictors, and predictions.

REFRESH STATUS

Unassigned Jobs:
(No jobs are waiting to be assigned.)

Jobs on Local Queue:

Name	Owner	Job Type	Number of Compounds	Number of Models	Time Created	Status	Cancel
sample	dpoz	PREDICTION	114	908	2013-09-06 08:15	Copying predictor	cancel

Jobs on LSF Queue:

Name	Owner	Job Type	Number of Compounds	Number of Models	Time Created	Status	Cancel
QSAR_365_SA-kNN_2	Rodaguayo	MODELING	365	4200	2013-08-05 01:27	Generating models (58%)	cancel

Prediction Results

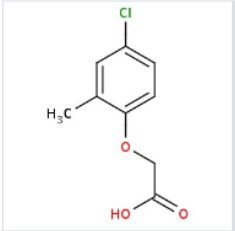
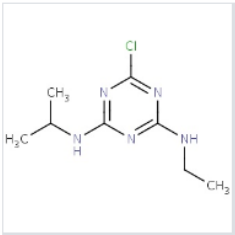
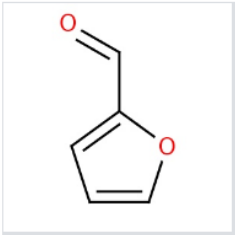
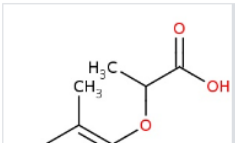
Prediction Name: Super_Fund_LD50_predict
Dataset Predicted: SuperFund
Predictors Used: RAT_ACUTE_LD50
Date Created: 2012-02-29 15:51
Similarity Cutoff: 1.0

Download This Prediction Result (CSV)
Back to Predictions

Prediction Values

Prediction Results

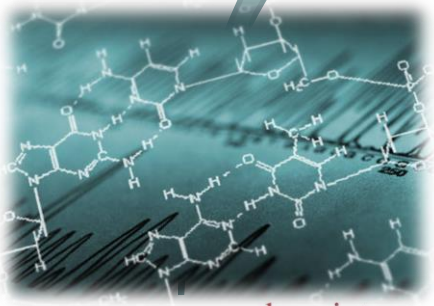
Go To Page: 1

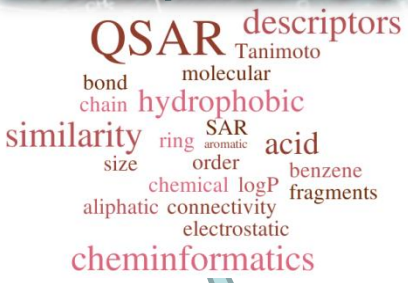
Compound ID	Structure	(RAT_ACUTE_LD50) Prediction	(RAT_ACUTE_LD50) Number of Predicting Models / Total Models
2-Methyl-4-chlorophenoxyacetic_acid__OB_MCPA_CB_		2.444 ± 0.139	568 / 568
Atrazine		2.333 ± 0.113	568 / 568
Furfural		2.135 ± 0.318	470 / 568
Methylchlorophenoxypropionic_acid__OB_MCPP_CB_		2.452 ± 0.129	568 / 568

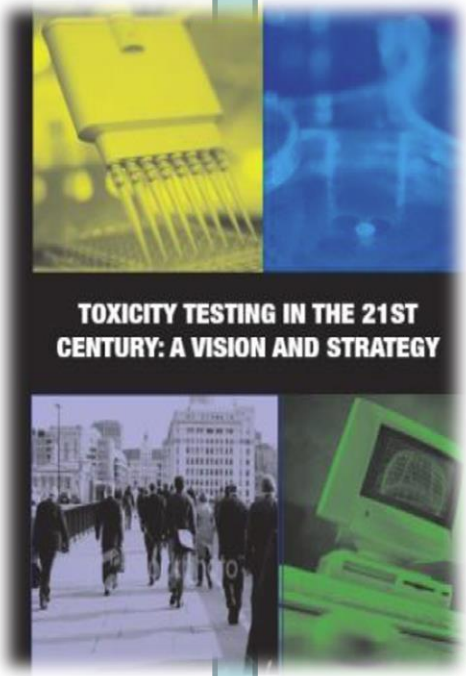
Integration of chemical descriptors and biological data streams to improve model accuracy and interpretability ("modeling with descriptors")

Cheminformatics

Over multiple chemicals

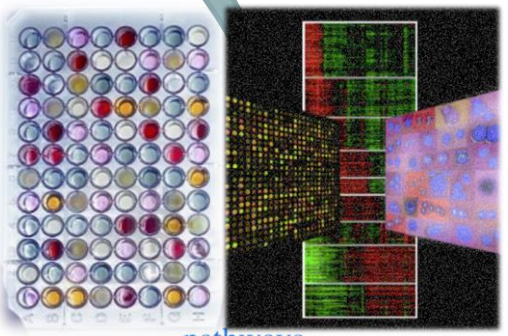


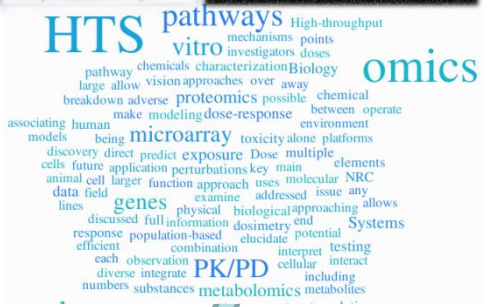




Bioinformatics

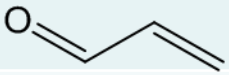
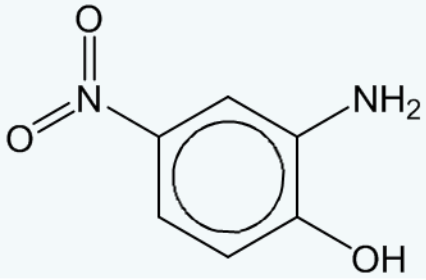
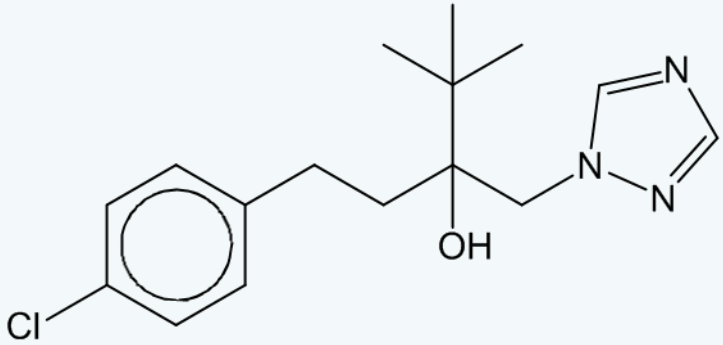
Over multiple biological assays





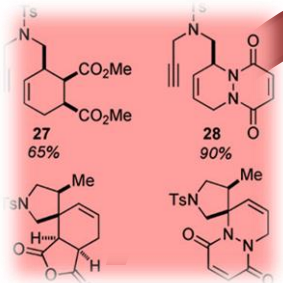
Human Toxicity

QSAR Table – biological (e.g., qHTS, gene expression, etc) descriptors

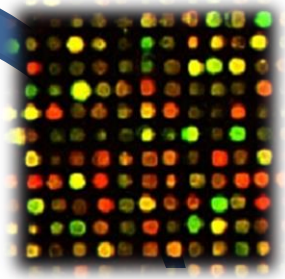
			Descriptor #: *1*	*2*	*…*	*182*
ID	**Name**	**Structure**	**3T3 9.2mkM**	**3T3 21mkM**	**…**	**SHSY 92mkM**
1	Acrolein		0	0	…	-92
2	2-Amino-4-nitrophenol		0	-22	…	0
…	…	…	…	…	…	…
369	Tebuco-nazole		-21	-24	…	-18

QSAR models < Hybrid models < Toxicogenomics models

Data source: TGP2 Toxicogenomics Informatics Project in Japan



Chemical descriptors



Toxicogenomics expression
(24h)

304 Dragon descriptors

Hybrid models
68- 75% BAcc

2,923 genes

Rank by differential expression

Top 400 genes

Top 100 genes

Top 30 genes

Top 4 genes

QSAR models
55-61% BAcc

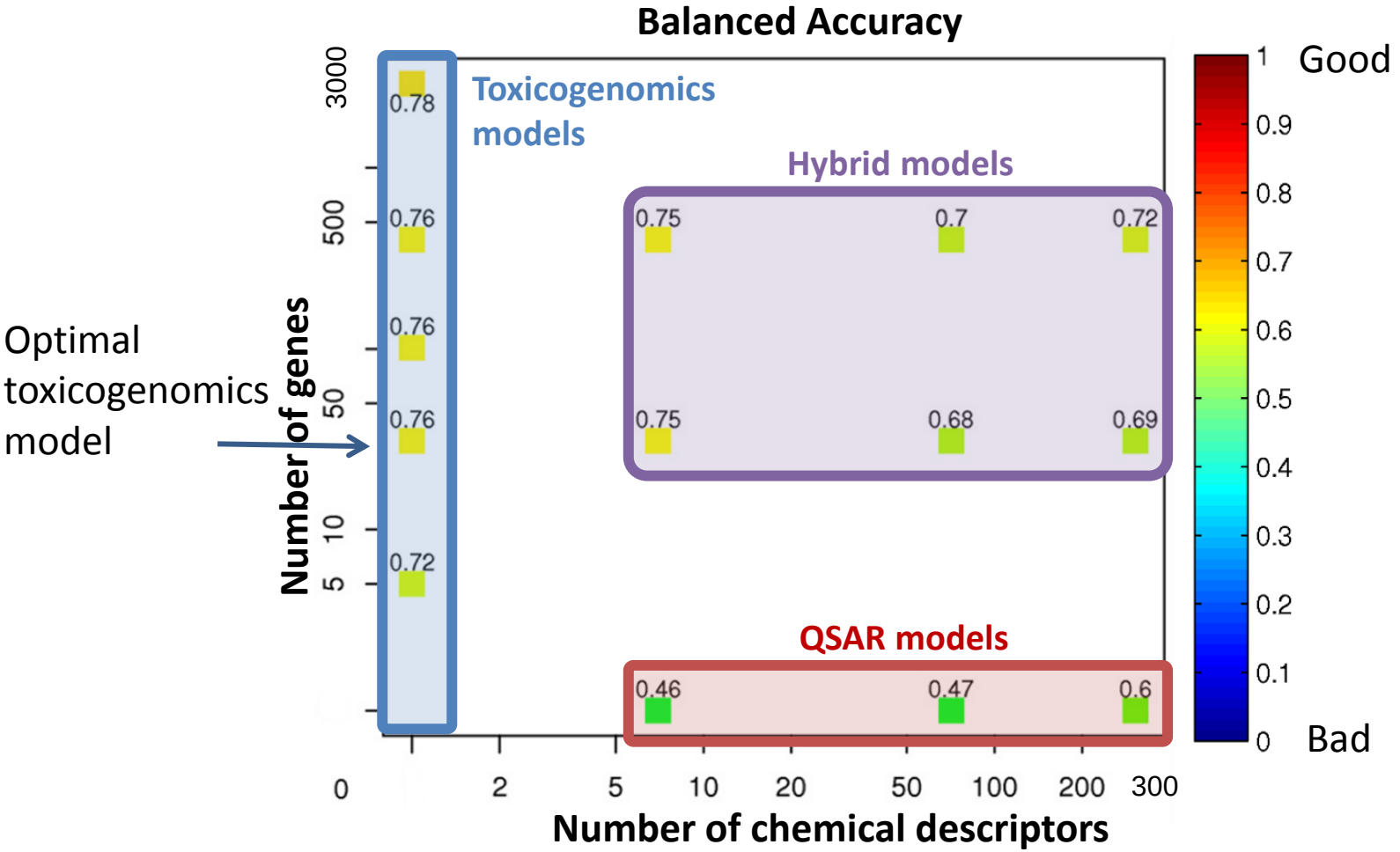
Toxicogenomics models
69-78% BAcc

Hepatotoxicity
(28 day)

4 classification methods
(RF, SVM, kNN, DWD)

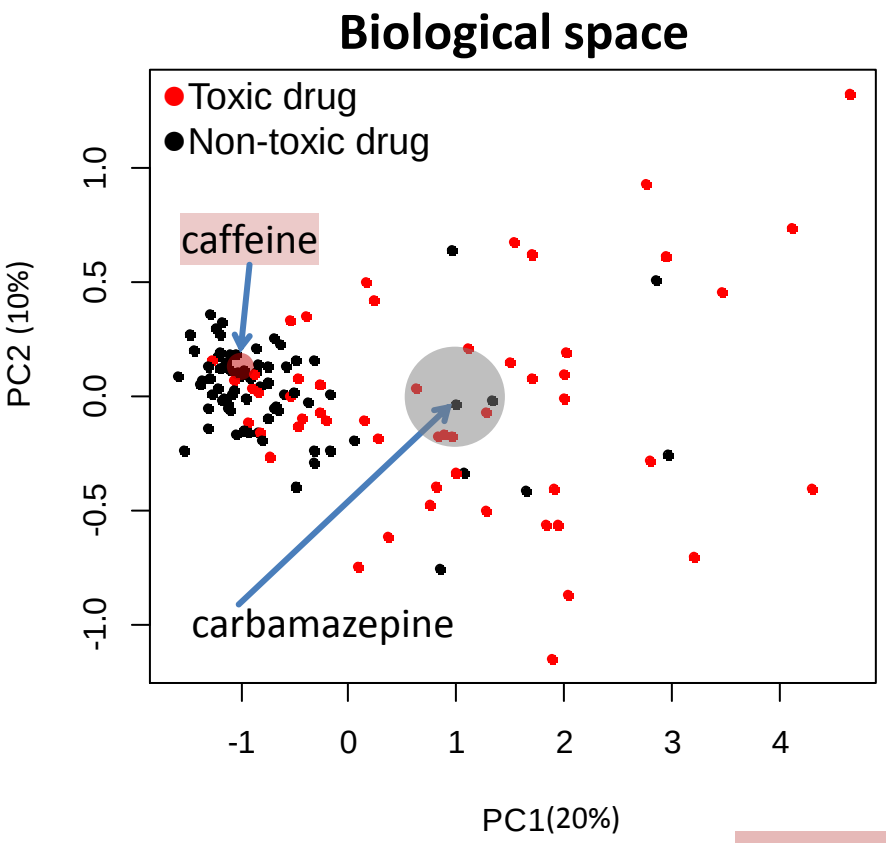
Low et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262

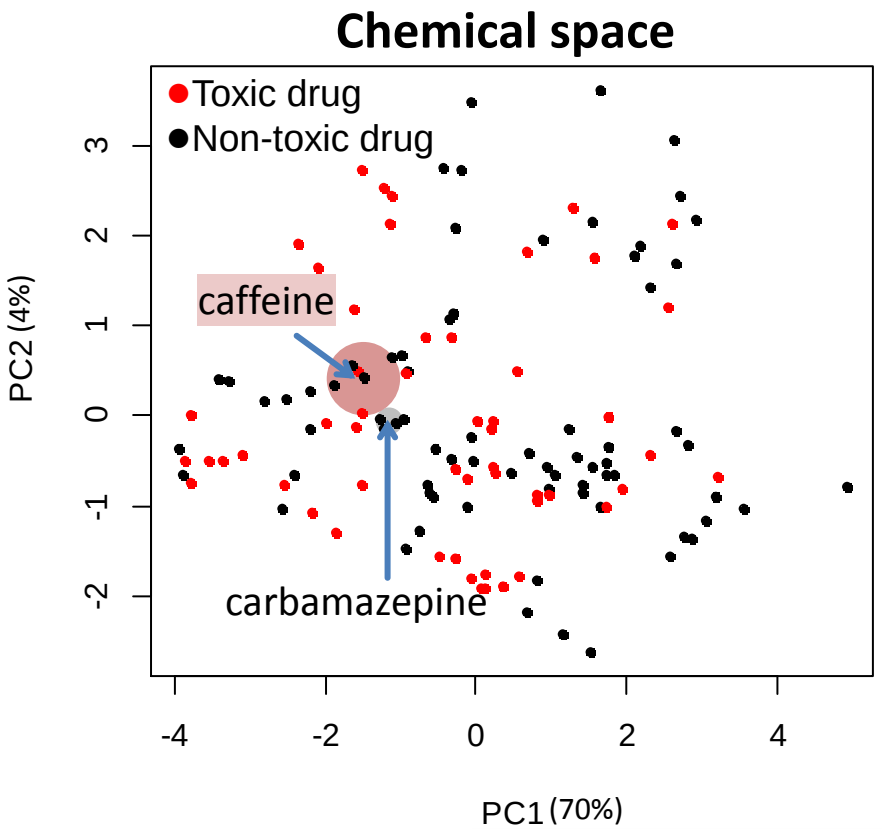
How predictivity varied with number of genes and number of chemical descriptors



Low et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262

Problem: Conflicting predictions by QSAR and toxicogenomics models





Carbamazepine

☒ Distant biological neighbors

☑ Close chemical neighbors

=> QSAR works better

Caffeine

☑ Close biological neighbors

☒ Distant chemical neighbors

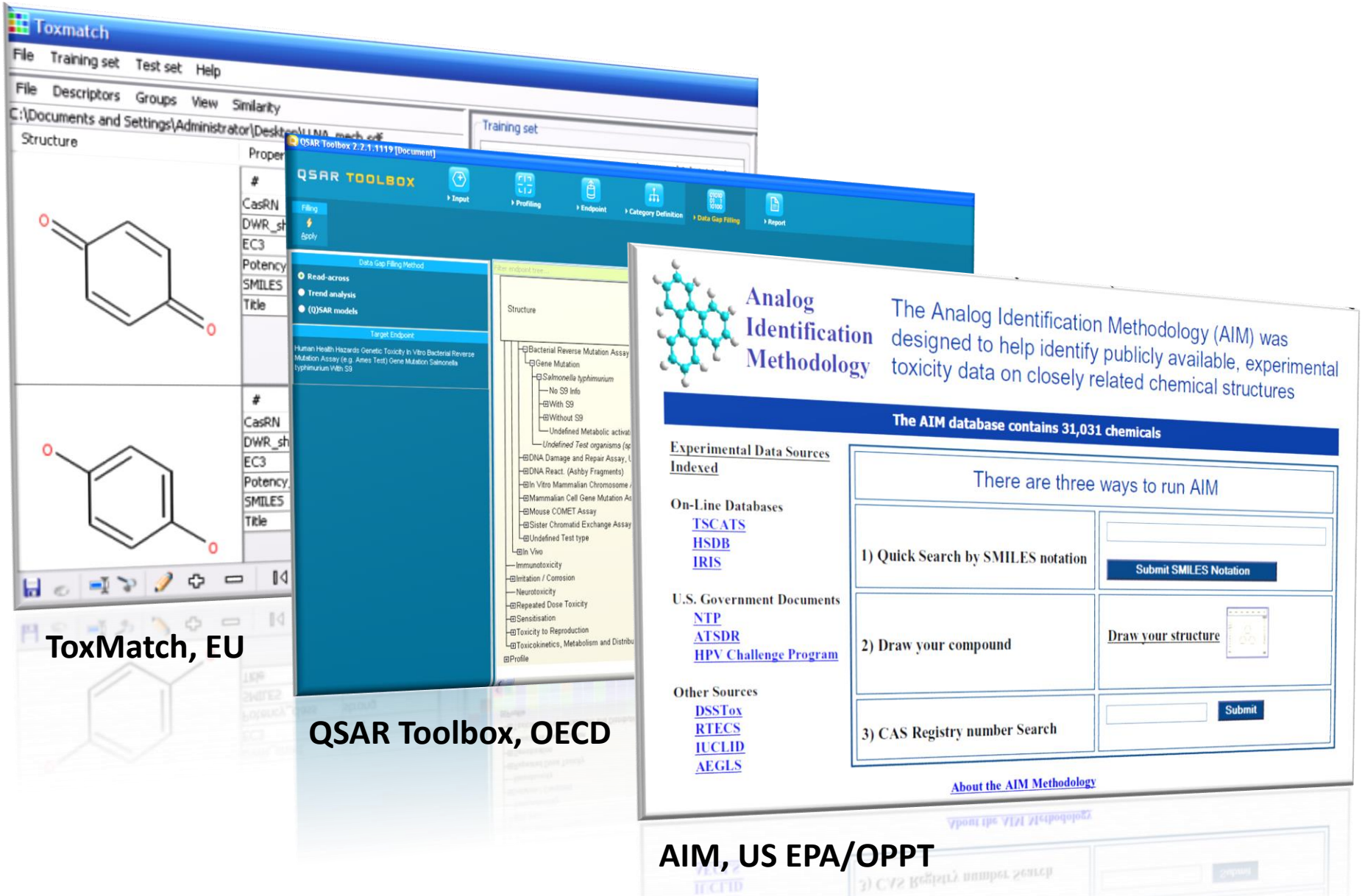
=> TGx works better

Solution:

Learn from both sets of neighbors

Learning from similar compounds

Traditional read-across predicts toxicity from chemically similar neighbors



ToxMatch, EU

QSAR Toolbox, OECD

AIM, US EPA/OPPT

Chemical-biological read-across (CBRA) learns from both sets of neighbors

Predicted toxicity =similarity-weighted average of toxicity values = $\frac{\sum_{i=1}^{k_{bio}} S_i \cdot A_i + \sum_{j=1}^{k_{chem}} S_j \cdot A_j}{\sum_{i=1}^{k_{bio}} S_i + \sum_{j=1}^{k_{chem}} S_j}$

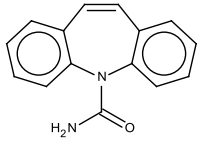
conflicting predictions

incorrectly predicted as toxic

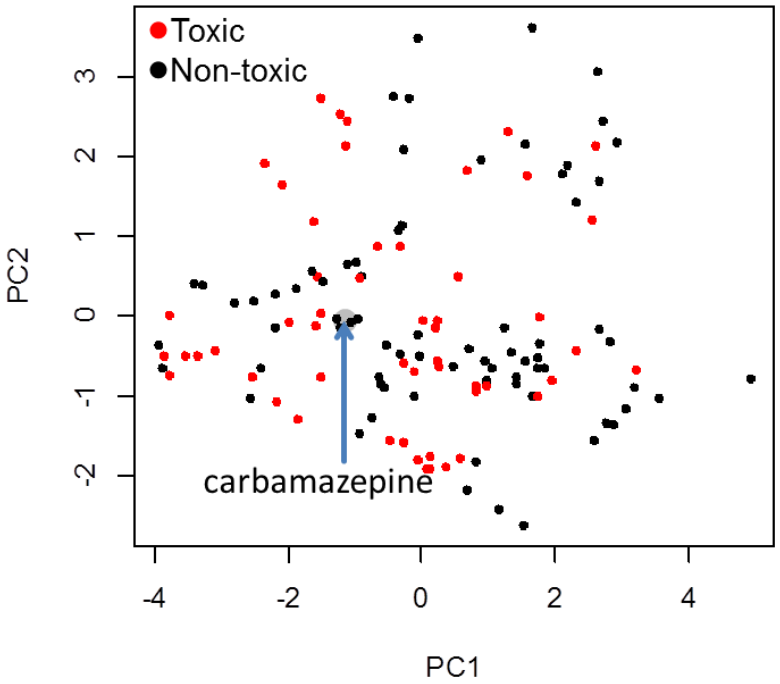
CARBAMAZEPINE

Non-toxic
Predicted as Non-toxic
Predicted toxicity=-0.099

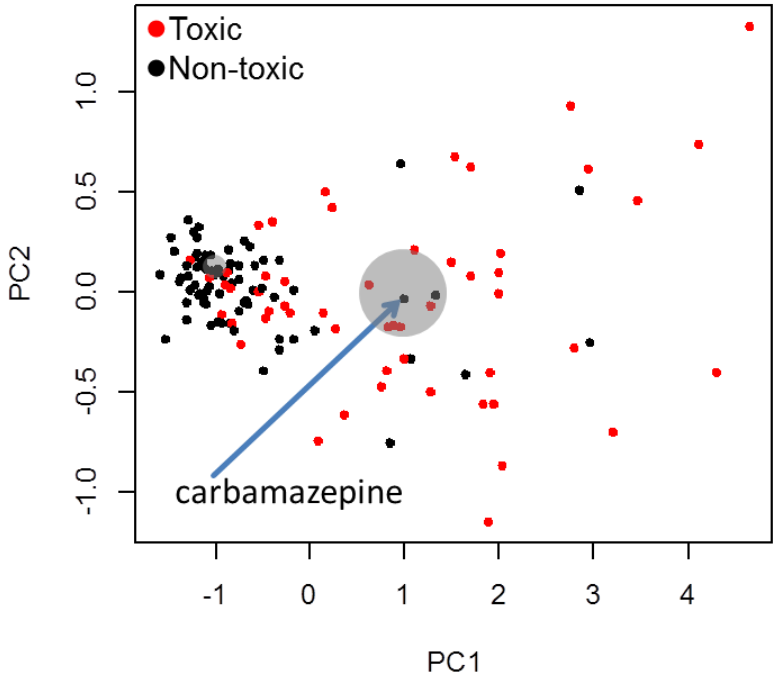
correctly predicted as nontoxic



Close chemical neighbors

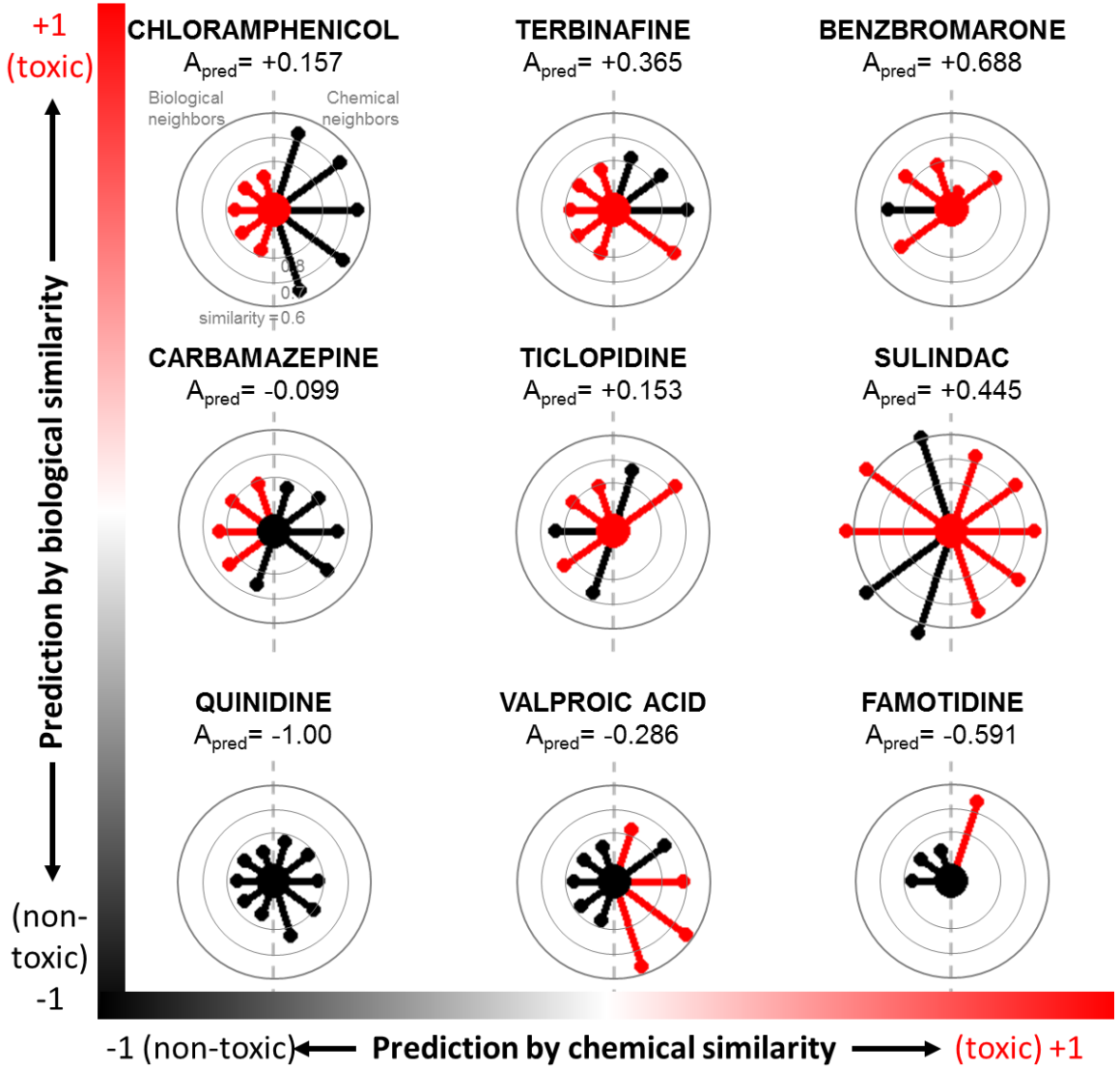


Distant biological neighbors



Low et al. (2013) *Chem. Res. Toxicol*. In press.

CBRA allows visual comparison of multiple compounds



Low et al. (2013) *Chem. Res. Toxicol.* In press.

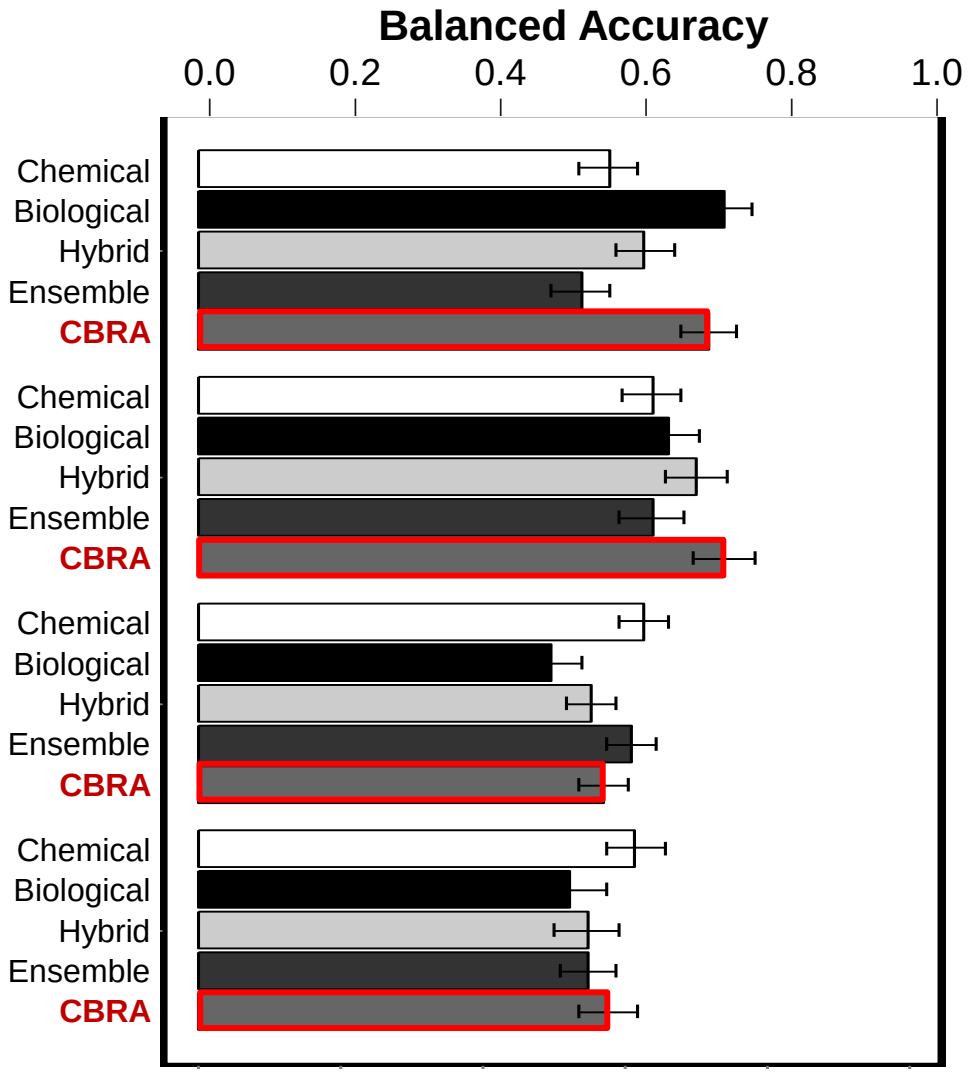
Results: CBRA consistently among the best models in 4 benchmark data sets

Rat Hepatotoxicity
127 compounds
85 genes

Rat Hepatocarcinogenicity
132 compounds
200 genes

Mutagenicity (Ames Test)
185 compounds
148 cytotoxicity assays

Rat Acute Toxicity (Oral LD_{50})
122 compounds
148 cytotoxicity assays

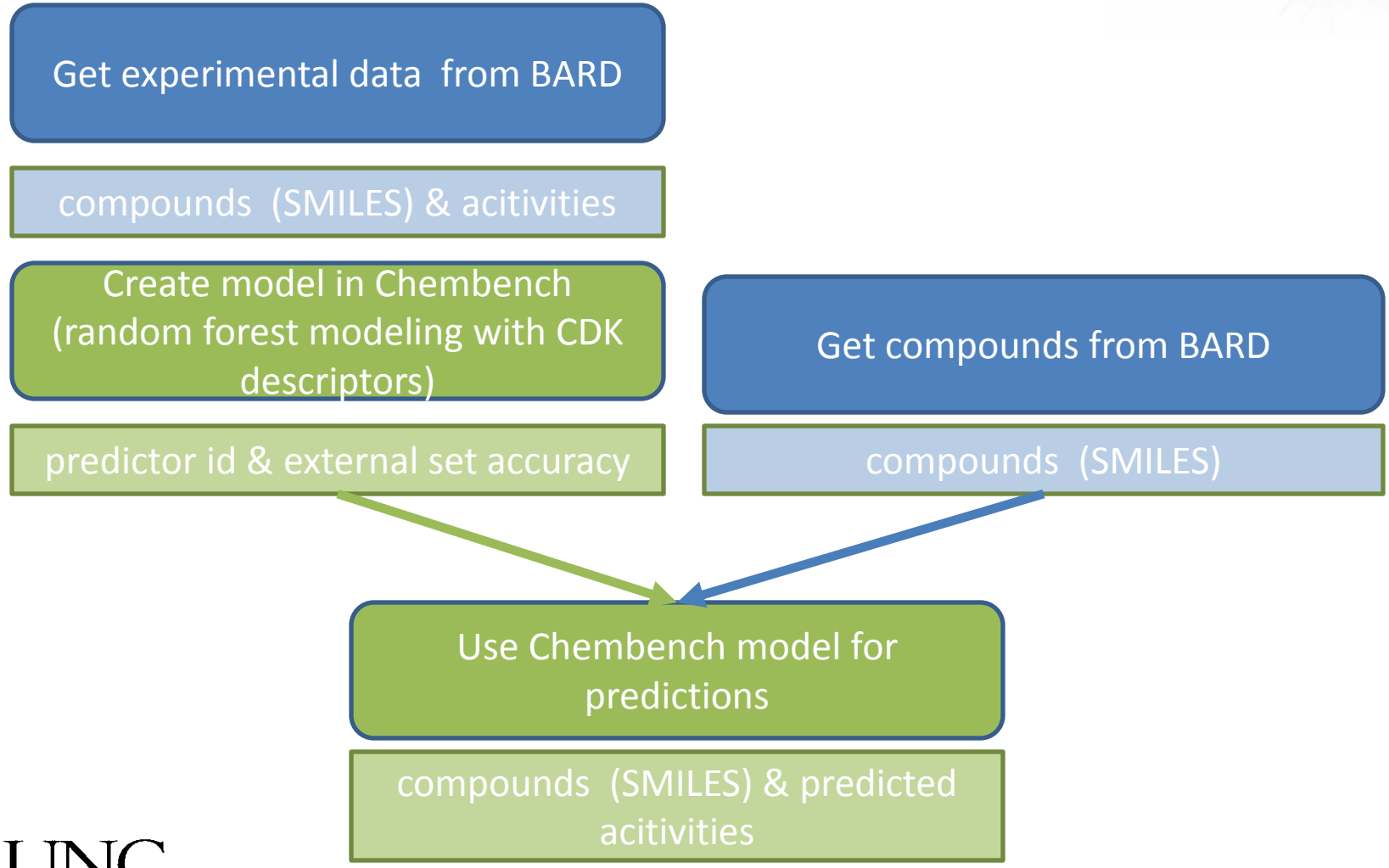


Low et al. (2013) *Chem. Res. Toxicol*. In press.

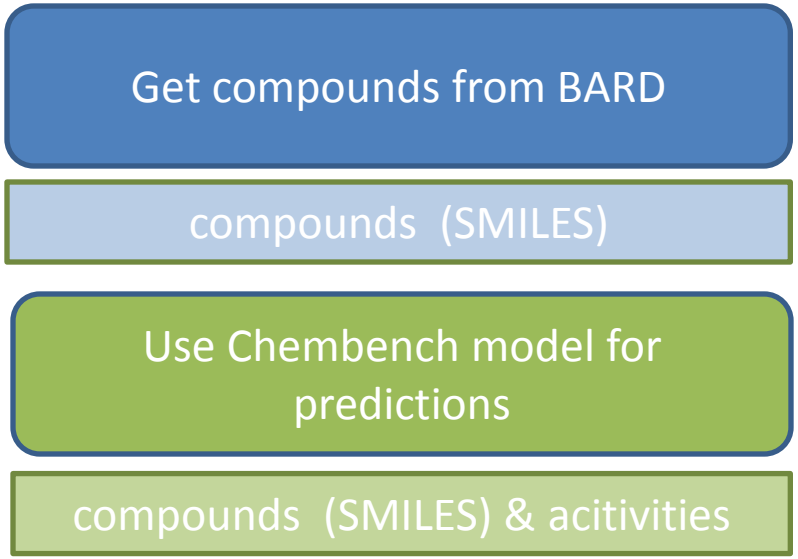
Chembench BARD Plugin
(under development)

- Take advantage of Chembench's
  - well defined workflow
  - publicly available models
- Complement BARD as data modeling tool
- Three types of use
  - Create a model from BARD's data
  - Run a virtual screening of a BARD dataset
  - Run a prediction on a single compound or any external library
- Predictions/virtual screenings can be run using
  - A predictor you have built ("private")
  - Publicly available predictors

Creating and using a model



Using a public model



The Laboratory for Molecular Modeling

Principal Investigator
Alexander Tropsha

Research Professors
Alexander Golbraikh, **Denis Fourches**, **Eugene Muratov**

Collaborators
Bryan Roth,
Ivan Rusyn
Nikolay Dokholyan

Postdoctoral Fellows
Aleck Sedykh,
Regina Politi

Graduate Research Assistants
Andrew Fant,
Stephen Bush,
Yen Low
Mary La

Adjunct Members
Weifan Zheng, Shubin Liu

MAJOR FUNDING

NIH
- R01-GM66940
- R01-GM068665

NSF
- ABI 9179-1165

EPA (STAR awards)
- RD832720
- RD833825
- RD834999

ONR